EX PARTE OR LATE FILED

BELLSOUTH

Kathleen B. Levitz
Vice President-Federal Regulatory

Suite 900
1133-21st Street, N.W.
Washington, D.C. 20036-3351
202 463-4113
Fax: 202 463-4198
Internet: levitz.kathleen@bsc.bls.com

February 18, 1999

RECEIVED

FEB 18 1999

FEDERAL COMMUNICATIONS COMMISSION
OFFICE OF THE SECRETARY

Ms. Magalie Roman Salas
Secretary
Federal Communications Commission
The Portals
445 12th St. S.W.
Washington, D.C. 20554

Re: Written Ex Parte in CC Docket No. 98-56 and CC Docket No. 98-121

Dear Ms. Salas:

This is to inform you that BellSouth Corporation has made a written ex parte to Dr. Daniel Shiman of the Common Carrier Bureau's Policy and Program Planning Division. That ex parte consists of a copy of "Controlling Bias in Observational Studies: A Review." This information has been submitted in response to Dr. Shiman's request.

Pursuant to Section 1.1206(a)(1) of the Commission's rules, I am filing two copies of this notice and that written ex parte presentation in both the dockets identified above. Please associate this notification with the record in both those proceedings.

Sincerely,

Kathleen B Levitz

Kathleen B. Levitz
Vice President- Federal Regulatory

Attachment

cc: Daniel Shiman (w/o attachment)
Florence Setzer (w/o attachment)
Alex Belinfante (w/o attachment)
Jake Jennings (w/o attachment)
Andre Rausch (w/o attachment)

BELLSOUTH

Kathleen B. Levitz
Vice President-Federal Regulatory

Suite 900
1133-21st Street, N W
Washington, D C 20036-3351
202 463-4113
Fax 202 463-4198
Internet: levitz.kathleen@bsc.bls.com

February 18, 1999

RECEIVED

FEB 18 1999

FEDERAL COMMUNICATIONS COMMISSION
OFFICE OF THE SECRETARY

Dr. Daniel Shiman
Policy and Program Planning Division
Common Carrier Bureau
Federal Communications Commission
1919 M Street, N.W.
Washington, D.C. 20554

Written Ex Parte in CC Docket No. 98-121 and
CC Docket No. 98-56

Dear Dr. Shiman:

Attached is a copy of the Cochran and Rubin paper "Controlling Bias in Observational Studies: A Review," which you have requested we send to you. If after reviewing it you conclude that you need additional information, please call me at (202) 463-4113.

In compliance with Section 1.1206(a)(1) of the Commission's rules, I have today filed with the Secretary of the Commission two copies of this written ex parte presentation for both CC Docket No. 98-56 and CC Docket No. 98-121 and requested that it be associated with the record of both dockets.

Sincerely,

Kathleen B. Levitz

Attachment

cc: Florence Setzer, Whitey Thayer, Alex Belinfante, Jake Jennings, Andre Rausch

CONTROLLING BIAS IN OBSERVATIONAL STUDIES : A REVIEW[1]

By WILLIAM G. COCHRAN[2]
Harvard University
and
DONALD B. RUBIN[3]
Educational Testing Service and *Princeton University*

SUMMARY. This paper reviews work on the effectiveness of different methods of matched sampling and statistical adjustment, alone and in combination, in reducing bias due to confounding x-variables when comparing two populations. The adjustment methods were linear regression adjustment for x continuous and direct standardization for x categorical.

With x continuous, the range of situations examined included linear relations between y and x, parallel and non-parallel, monotonic non-linear parallel relations, equal and unequal variances of x, and the presence of errors of measurement in x.

The percent of initial bias $E(\bar{y}_1 - \bar{y}_2)$ that was removed was used as the criterion. Overall, linear regression adjustment on random samples appeared superior to the matching methods, with linear regression adjustment on matched samples the most robust method. Several different approaches were suggested for the case of multivariate x, on which little or no work has been done.

1. Introduction

An observational study differs from an experiment in that the random assignment of treatments (i.e. agents, programs, procedures) to units is absent. As has been pointed out by many writers since Fisher (1925), this randomization is a powerful tool in that many systematic sources of bias are made random. If randomization is absent, it is virtually impossible in many practical circumstances to be convinced that the estimates of the effects of treatments are in fact unbiased. This follows because other variables that affect the dependent variable besides the treatment may be differently distributed across treatment groups, and thus any estimate of the treatment is confounded by these extraneous x-variables.

Given the choice between an observational study and an essentially equivalent randomized experiment one would prefer the experiment. Thus in the Report of the President's Commission on Federal Statistics (1971), Light, Mosteller, and Winokur urge greater efforts to use randomized studies in evaluating public programs and in social experimentation, despite the practical difficulties. Often however, random assignment of treatments to units is not feasible, as in the studies of the effects of smoking on health, complications of pregnancy on children, or long-term exposure to

[1]Requests for reprints should be addressed to Donald B. Rubin, Educational Testing Service, Princeton, New Jersey 08540, U.S.A.

[2]Supported by a contract with the Office of Naval Research, Navy Department.

[3]Partially supported by the U.S. Office of Education under contract OEC-0-71-3715.

doses of radiation on uranium mine workers. Also, as in these examples, one might have to wait many years for the results of an experiment while relevant observational data might be at hand. Hence, although inferior to an equivalent experiment, an observational study may be superior to or useful in conjunction with a marginally relevant experiment (e.g. one on the long-term effects of radiation on white rats). In addition, the analysis of data from observational studies can be useful in isolating those treatments that appear to be successful and thus worth further investigation by experimentation, as when studying special teaching methods for underprivileged children.

In dealing with the presence of confounding variables, a basic step in planning an observational study is to list the major confounding variables, design the study to record them, and find some method of removing or reducing the biases that they may cause. In addition, it is useful to speculate about the size and direction of any remaining bias when summarizing the evidence on any differential effects of the treatments.

There are two principal strategies for reducing bias in observational studies. In matching or matched sampling, the samples are drawn from the populations in such a way that the distributions of the confounding variables are similar in some respects in the samples. Alternatively, random samples may be drawn, the estimates of the treatment being adjusted by means of a model relating the dependent variable y to the confounding variable x. When y and x are continuous, this model usually involves the regression of y on x. A third strategy is to control bias due to the x-variables by both matched sampling and statistical adjustment. Notice that the statistical adjustment is performed after all the data are collected, while matched sampling can take place before the dependent variable is recorded.

This paper reviews work on the effectiveness of matching and statistical adjustments in reducing bias in a dependent variable y and two populations P_1 and P_2 defined by exposure to two treatments. Here, the objective is to estimate the difference $(\tau_1-\tau_2)$ between the average effects of the treatments on y.

Section 2 reviews work on the ability of linear regression adjustment and three matching methods to reduce the bias due to x in the simplest case when both y and x are continuous, there are parallel linear regressions in both populations, and x is the only confounding variable. Section 3 considers complications to this simple case : non-parallel regressions, non-linear regressions, errors of measurement in x, and the effect of an omitted confounding variable. Section 4 extends the above cases to include x categorical or made categorical (e.g. low, medium, high). Section 5 presents some multivariate x results which are simple generalizations of the univariate x results. Section 6 considers some multivariate extensions of matching methods. A brief summary of the results and indications for further research are given in Section 7.

2. y, x CONTINUOUS : UNIVARIATE PARALLEL LINEAR REGRESSIONS

2.1. *The model.* We begin with the simple case when y and a univariate x are both continuous, and the regressions of y on x are linear and parallel in both populations. For the j-th observation from population i, the model may be written

$$y_{ij} = \mu_i + \beta(x_{ij} - \eta_i) + e_{ij} \qquad \text{... (2.1.1)}$$

with

$$E(e_{ij} \mid x_{ij}) = 0, \quad E(e_{ij}^2 \mid x_{ij}) = \sigma_i^2$$

where μ_i and η_i are the means of y and x respectively in population i, where $\eta_1 > \eta_2$ without loss of generality. Thus the regressions of y on x differ by the constant

$$E(y_{1j} - y_{2j} \mid x_{1j} = x_{2j}) = (\mu_1 - \mu_2) - \beta(\eta_1 - \eta_2). \qquad \text{... (2.1.2)}$$

If x is the only variable (besides the treatment) that affects y *and* whose distribution differs in the two populations, (2.1.2) equals the difference in the average effects of the treatments, $\tau_1 - \tau_2$. Thus, in this case, the treatment difference in (2.1.2) is constant at any level of x.

From (2.1.1) it follows that conditionally on the values of x_{ij} in samples chosen either randomly or solely on x,

$$\begin{aligned} E_c(\bar{y}_1 - \bar{y}_2) &= (\mu_1 - \mu_2) + \beta(\bar{x}_1 - \eta_1) - \beta(\bar{x}_2 - \eta_2) \\ &= \tau_1 - \tau_2 + \beta(\bar{x}_1 - \bar{x}_2). \end{aligned} \qquad \text{... (2.1.3)}$$

Letting E_r be the expectation over the distribution of variables in random samples,

$$E_r(\bar{y}_1 - \bar{y}_2) = \mu_1 - \mu_2 = \tau_1 - \tau_2 + \beta(\eta_1 - \eta_2) \qquad \text{... (2.1.4)}$$

so that the expected bias in $(\bar{y}_1 - \bar{y}_2)$ from random samples is $\beta(\eta_1 - \eta_2)$.

2.2. *Linear regression adjustment.* Since from (2.1.3) $\bar{y}_1 - \bar{y}_2$ is conditionally biased by an amount $\beta(\bar{x}_1 - \bar{x}_2)$ in random and matched samples, it is reasonable to adjust $\bar{y}_1 - \bar{y}_2$ by subtracting an estimate of the bias. The adjusted estimate would then be

$$\hat{\tau}_1 - \hat{\tau}_2 = (\bar{y}_1 - \bar{y}_2) - \hat{\beta}(\bar{x}_1 - \bar{x}_2).$$

In practice, $\hat{\beta}$ is most commonly estimated from the pooled within-sample regressions. With this model, however, $E_c(\hat{\beta}) = \beta$ either for the pooled $\hat{\beta}$ or for $\hat{\beta}$ estimated from sample 1 or sample 2 alone. From (2.1.3) for any of these $\hat{\beta}$,

$$E_c(\hat{\tau}_1 - \hat{\tau}_2) = \mu_1 - \mu_2 - \beta(\eta_1 - \eta_2) = \tau_1 - \tau_2.$$

For this model, the regression adjustment removes all the bias either for random samples or for matched samples selected solely using x.

Before using the regression adjusted estimate, the investigator should satisfy himself that the regressions of y on x in the two populations appear linear and parallel. Standard methods of fitting higher order terms in x and separate β's in the two samples are appropriate for helping to answer this question.

2.3. *Caliper matching.* In order to construct matched samples of size n, the investigator needs initial reservoirs of data of sizes $r_1 n$, $r_2 n$ from which to seek matches, where $r_i \geq 1$ with at least one $r_i > 1$. The work to be reported here is for the case $r_1 = 1$, in which there is a random sample of size n from population 1 to which the sample from population 2 is to be matched from a reservoir of size rn $(r > 1)$. This case is appropriate in studies in which population 1 is of primary interest, population 2 being a control population (untreated or with a standard treatment) with a larger reservoir from which a sample matched to sample 1 is drawn. The case of only one reservoir is a fairly severe test for matching since it is easier to obtain close matches with reservoirs from both populations.

With a random sample from population 1 and some kind of matched sample from population 2 chosen using x, relation (2.1.3) gives the expected bias of matched samples as

$$E_m(\bar{y}_1 - \bar{y}_2) = (\tau_1 - \tau_2) + \beta(\eta_1 - E_m(\bar{x}_2)) \quad \ldots \quad (2.3.1)$$

where E_m is the expectation over the distribution of variables in samples from population 2 matched on x.

The criterion to be used in judging the effectiveness of matching will be the percentage reduction in bias. From (2.1.4) and (2.3.1) this is

$$\theta = (100)\,\frac{E_m(\bar{x}_2) - \eta_2}{\eta_1 - \eta_2} \quad \ldots \quad (2.3.2)$$

We note that with this model the percentage reduction in expected bias of $(\bar{y}_1 - \bar{y}_2)$ equals that in $(\bar{x}_1 - \bar{x}_2)$.

As a measure of the amount of initial bias in x when appraising methods of matching or adjustment, we chose the quantity

$$B = (\eta_1 - \eta_2) \Big/ \left(\frac{\sigma_1^2 + \sigma_2^2}{2}\right)^{1/2}$$

and examined values of B in the range $(0, 1)$. A value of $B = 1$ is considered large. With this bias, the difference $(\bar{x}_1 - \bar{x}_2)$ has about a 90% chance of being detected as significant (5% level) in random samples of 25 when σ_1^2, σ_2^2 are not too unequal. The values of σ_1^2/σ_2^2 studied were $\frac{1}{2}$, 1, 2.

The first method of matching investigated, often used with x continuous, is paired caliper matching. Each x_{1j} has a partner x_{2j} such that

$$|x_{1j} - x_{2j}| \leq a.$$

This method is attractive from two points of view. Although we are assuming at present a *linear* regression of y on x, it is clear that a tight caliper matching should remove nearly all the bias in $(\bar{y}_1 - \bar{y}_2)$ under any smooth regression, linear or non-linear, that is the same in both populations. Secondly, at first sight this method provides convenient data for investigating how $E_c(y_{1j} - y_{2j})$ varies with x, since x is close to constant for any single pair.

In presenting results on the percent reductions in bias for x normal (Table 2.3.1), we have taken

$$c = a\sqrt{(\sigma_1^2+\sigma_2^2)/2}$$

where $a = 0.2(0.2)1.0$. Strictly, the results hold for $B < 0.5$ but for B between 0.5 and 1, the percent reductions are only about 1 to $1\frac{1}{2}$% lower than the figures shown.

TABLE 2.3.1. PERCENT REDUCTION IN BIAS OF x FOR CALIPER MATCHING TO WITHIN $\pm a\sqrt{(\sigma_1^2+\sigma_2^2)/2}$ WITH x NORMAL

a	$\sigma_1^2/\sigma_2^2 = \frac{1}{2}$	$\sigma_1^2/\sigma_2^2 = 1$	$\sigma_1^2/\sigma_2^2 = 2$
0.2	.99	.99	.98
0.4	.96	.95	.93
0.6	.91	.89	.86
0.8	.86	.82	.77
1.0	.79	.74	.69

A tight matching ($a = 0.2$) removes practically all the bias, while a loose matching ($a = 1.0$) removes around 75%. The ratio σ_1^2/σ_2^2 has a minor effect, although performance is somewhat poorer as σ_1^2/σ_2^2 increases.

A disadvantage of caliper matching in practical use is that unless r is quite large there is a non-negligible probability that some of the desired n matches are not found in the reservoir. Nothing seems to be known about the distribution of the number of matches found as a function of r, a, $(\eta_1-\eta_2)$ and σ_1^2/σ_2^2. We have not investigated the consequences of incomplete matching as often results in practice. Thus we have no help to give the investigator in estimating the reservoir size needed and the probable percent success in finding caliper matches.

2.4. '*Nearest available' matching*. This disadvantage is avoided by a method, (Rubin, 1973a), in which all n pair matches are easily formed by computer. The n values of x from sample 1 and the rn values from reservoir 2 are entered in the computer. In one variant of the method, the sample 1 values of x are first arranged in random order from x_{11} to x_{1n}. Starting with x_{11}, the computer selects the value x_{21} in reservoir 2 nearest to x_{11} and lays this pair aside. The computer next seeks a 'nearest available' partner for x_{12} from the $(rn-1)$ remaining in reservoir 2, and so on, so that n matches are always found although the value of a is not controlled.

Two other variants of this 'nearest available' method were examined. In these, the members of sample 1 were (i) first ranked from highest to lowest, (ii) first ranked from lowest to highest, before seeking matches from the ranked samples. For $\eta_1 > \eta_2$, Monte Carlo results with x normal showed that for the percent reductions θ in bias of $(\bar{x}_1-\bar{x}_2)$, $\theta_{LH} > \theta_{ran} > \theta_{HL}$. If, however, the quality of the matches is

3

judged by the average MSE within pairs, $E_m(x_{1j}-x_{2j})^2$, the order of performance was opposite : $\text{MSE}_{HL} < \text{MSE}_{ran} < \text{MSE}_{LH}$. Both sets of results have rational explanations. The differences in performance were usually small. On balance, random ordering is a reasonable compromise as well as quickest for the computer.

For random ordering, Table 2.4.1 shows the percent reductions in bias of $(\bar{x}_1-\bar{x}_2)$ and hence of $(\bar{y}_1-\bar{y}_2)$ for $r = 2, 3, 4$, $n = 25, 50$ and different combinations of the initial bias B and the σ_1^2/σ_2^2 ratio. Results for $n = 100$ (not shown) differ by at most one or two percentage points from those for $n = 50$, suggesting that the $n = 50$ results hold also for $n > 50$. With this method, the percent reduction in bias decreases steadily as the bias B increases from 1/4 to 1, so that results are given separately for the four values of B.

As regards the effect of σ_1^2/σ_2^2, matching does best when $\sigma_1^2/\sigma_2^2 = \frac{1}{2}$ and worst when $\sigma_1^2/\sigma_2^2 = 2$. This is not surprising. Since $\eta_1 > \eta_2$ the high values of sample 1 (the ones most likely to cause residual bias) will receive less biased partners when $\sigma_2^2 > \sigma_1^2$.

The investigator planning to use 'nearest available' matching can estimate B and σ_1^2/σ_2^2 from the initial data on x. Knowing the value of r, he can estimate the expected percent reduction in bias under a linear regression from Table 2.4.1.

TABLE 2.4.1. PERCENT REDUCTION IN BIAS FOR RANDOM ORDER, NEAREST AVAILABLE MATCHING; x NORMAL

		$\sigma_1^2/\sigma_2^2 = \frac{1}{2}$				$\sigma_1^2/\sigma_2^2 = 1$				$\sigma_1^2/\sigma_2^2 = 2$			
	$r \diagdown B =$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	1	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	1	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	1
$n = 25$	2	97	94	89	80	87	82	75	66	63	60	56	48
	3	99	98	97	93	94	91	86	81	77	72	67	61
	4	99	99	99	97	96	95	92	88	81	79	76	68
$n = 50$	2	99	98	93	84	92	87	78	69	66	59	53	51
	3	100	99	99	97	96	96	91	84	79	75	69	63
	4	100	100	100	99	98	97	94	89	86	81	75	71

A measure has also been constructed (Rubin, 1973a) of the closeness or quality of the individual pair matches. If pairing were entirely at random, we would have

$$E_m(x_{1j}-x_{2j})^2 = (\sigma_1^2+\sigma_2^2)+(\eta_1-\eta_2)^2$$
$$= (\sigma_1^2+\sigma_2^2)(1+B^2/2).$$

Consequently the quantity

$$100E_m(x_{1j}-x_{2j})^2/(\sigma_1^2+\sigma_2^2)(1+B^2/2)$$

was chosen as the measure. Since results vary little with n, only those for $n = 50$ are shown in Table 2.4.2.

TABLE 2.4.2. VALUES OF $100E_m(x_{1j}-x_{2j})^2/(\sigma_1^2+\sigma_2^2)(1+B^2/2)$ FOR NEAREST AVAILABLE RANDOM ORDER MATCHING WITH x NORMAL

	$\sigma_1^2/\sigma_2^2 = \frac{1}{2}$				$\sigma_1^2/\sigma_2^2 = 1$				$\sigma_1^2/\sigma_2^2 = 2$			
$r \backslash B =$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	1	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	1	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	1
2	0	1	3	8	1	2	5	15	7	12	20	26
3	0	0	0	1	0	1	3	6	4	8	12	18
4	0	0	0	0	0	1	2	4	3	5	9	12

Except for $\sigma_1^2/\sigma_2^2 = 2$ and $B > \frac{1}{2}$, random ordering gives good quality matches. In fact, since the computer program (Rubin, 1973a) for constructing the matched pairs is very speedy, the investigator can try random, high-low, and low-high ordering. By examining $(\bar{x}_1-\bar{x}_2)$ and $\Sigma(x_{1j}-x_{2j})^2/n$ for each method, he can select what appears to him the best of the three approaches.

2.5. *Mean matching.* For an investigator who is not interested in pair matching and is confident that the regression is linear, a mean-matching method which concentrates on making $|\bar{x}_1-\bar{x}_2|$ small has been discussed (Greenberg, 1953). The following simple computer method has been investigated (Rubin, 1973a). Calculate $\bar{x}_1$. Select, from reservoir 2, the x_{21} closest to $\bar{x}_1$, then the x_{22} such that $(x_{21}+x_{22})/2$ is closest to $\bar{x}_1$, and so on until n have been selected. For $n = 50$, Table 2.5.1 shows the percent reductions in bias obtained.

TABLE 2.5.1. PERCENT REDUCTION IN BIAS FOR MEAN MATCHING : x NORMAL

	$\sigma_1^2/\sigma_2^2 = \frac{1}{2}$				$\sigma_1^2/\sigma_2^2 = 1$				$\sigma_1^2/\sigma_2^2 = 2$			
$r \backslash B =$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	1	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	1	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	1
2	100	100	98	87	100	99	91	77	100	96	82	67
3	100	100	100	100	100	100	99	96	100	100	97	84
4	100	100	100	100	100	100	100	100	100	100	100	95

Except in a few difficult cases, particularly $B = 1$, this method of mean matching removes essentially all the bias. So far as we know, mean matching is seldom used, presumably because it relies heavily on the assumption that the regression is linear. With a monotone non-linear regression of y on x, one might speculate that mean matching should perform roughly as well as a linear regression adjustment on random samples. But with the regression adjustment, one can examine the relations between y and x in the two samples before deciding whether a linear or non-linear regression adjustment is appropriate, whereas with mean matching performed before y has been observed, one is committed to the assumption of linearity, at least when matching the samples

3. Complications

3.1. *Regressions linear but not parallel.* For $i = 1, 2$, the model becomes

$$y_{ij} = \mu_i+\beta_i(x_{ij}-\eta_i)+e_{ij}. \qquad \dots \quad (3.1.1)$$

It follows that for a given level of x,

$$E\{(y_{1j}-y_{2j})\,|\,x_{1j} = x_{2j} = x\} = \mu_1-\mu_2-\beta_1\eta_1+\beta_2\eta_2+(\beta_1-\beta_2)x. \qquad \dots \quad (3.1.2)$$

If this quantity is interpreted as measuring the difference in the effects of the two treatments for given x, this difference appears to have a linear regression on x. At this point the question arises whether a differential treatment effect with x is a reasonable interpretation or whether the $(\beta_1-\beta_2)$ difference is at least partly due to other characteristics (e.g., effect of omitted x-variables) in which the two populations differ. With samples from two populations treated differently, we do not see how this question can be settled on statistical evidence alone. With one study population P_1 and two control populations P_2, P_2' both subject to τ_2, a finding that $\hat{\beta}_2$ and $\hat{\beta}_2'$ agree closely but differ from $\hat{\beta}_1$ leans in favour of suggesting a differential effect of $(\tau_1-\tau_2)$.

As it happens, assuming x is the only confounding variable, this issue becomes less crucial if the goal is to estimate the average $(\tau_1-\tau_2)$ difference over population 1. From (3.1.2) this quantity is

$$E_1(\tau_1-\tau_2) = (\mu_1-\mu_2)-\beta_2(\eta_1-\eta_2). \qquad \dots \quad (3.1.3)$$

Since from random samples,

$$E_r(\bar{y}_1-\bar{y}_2) = \mu_1-\mu_2, \qquad \dots \quad (3.1.4)$$

the initial bias is $\beta_2(\eta_1-\eta_2)$. With samples matched to a random $\bar{x}_1$,

$$E_m(\bar{y}_1-\bar{y}_2) = \mu_1-\mu_2-\beta_2E_m(\bar{x}_2)+\beta_2\eta_2,$$

so that the reduction in bias is

$$E_r(\bar{y}_1-\bar{y}_2)-E_m(\bar{y}_1-\bar{y}_2) = \beta_2[E_m(\bar{x}_2)-\eta_2].$$

Hence the percent reduction in bias due to matching remains, as before,

$$100[E_m(\bar{x}_2)-\eta_2]/(\eta_1-\eta_2)$$

so that previous results for matching apply to non-parallel lines also with this estimand.

As regards regression adjustment, it follows from (3.1.3) and (3.1.4) that

$$E_r[(\bar{y}_1-\bar{y}_2)-\hat{\beta}_2(\bar{x}_1-\bar{x}_2)] = (\mu_1-\mu_2)-\beta_2(\eta_1-\eta_2) = E_1(\tau_1-\tau_2).$$

Consequently, in applying the regression adjustment to random samples, use of the regression coefficient calculated from sample 2 provides an unbiased estimate of the desired $E_1(\tau_1-\tau_2)$. This property was noted by Peters (1941), while Belson (1956) recommended the use of $\hat{\beta}_2$ in comparing listeners (P_1) with non-listeners (P_2) to a BBC television program designed to teach useful French words and phrases to prospective tourists.

With $E_1(\tau_1-\tau_2)$ as the objective, the standard use of the pooled $\hat{\beta}_p$ in the regression adjustment gives biased estimates, though Rubin (1970) has shown that 'nearest available' matching followed by regression adjustment greatly reduces this bias. With matched samples, the standard estimate of β, following the analysis of covariance in a two-way table, is $\hat{\beta}_d$, the sample regression of matched pair differences, $(y_{1j}-y_{2j})$ on $(x_{1j}-x_{2j})$. Curiously, the Monte Carlo computations show that use of $\hat{\beta}_p$ on matched samples performs better than use of $\hat{\beta}_d$ in this case.

If non-parallelism is interpreted as due to a $(\tau_1-\tau_2)$ difference varying linearly with x, the question whether $E_1(\tau_1-\tau_2)$ is the quantity to estimate deserves serious consideration. To take a practice sometimes followed in vital statistics, we might wish to estimate $(\tau_1-\tau_2)$ averaged over a standard population that has mean η_s differing from η_1 and η_2. The estimand becomes, from (3.1.2)

$$E_s(\tau_1-\tau_2) = \mu_1-\mu_2+\beta_1(\eta_s-\eta_1)-\beta_2(\eta_s-\eta_2).$$

From random samples, an unbiased regression estimate is

$$(\bar{y}_1-\bar{y}_2)+\hat{\beta}_1(\eta_s-\bar{x}_1)-\hat{\beta}_2(\eta_s-\bar{x}_2) \qquad \ldots \quad (3.1.5)$$

where $\hat{\beta}_1$ and $\hat{\beta}_2$ are the usual least squares estimates from the separate regressions in the two samples.

Alternatively, particularly if $\hat{\beta}_1$ and $\hat{\beta}_2$ differ substantially, no single average of $(\tau_1-\tau_2)$ may be of interest, but rather the values of $(\tau_1-\tau_2)$ at each of a range of values of x. As a guide in forming a judgement whether use of a single average difference is adequate for practical application, Rubin (1970) has suggested the following. Suppose that in the range of interest, x lies between x_L and x_H. From (3.1.2) the estimated difference in $(\tau_1-\tau_2)$ at these two extremes is

$$(\hat{\beta}_1-\hat{\beta}_2)(x_H-x_L). \qquad \ldots \quad (3.1.6)$$

From (3.1.5), the average $(\tau_1-\tau_2)$ over the range from x_L to x_H is estimated as

$$(\bar{y}_1-\bar{y}_2)+\hat{\beta}_1(\bar{x}-\bar{x}_1)-\hat{\beta}_2(\bar{x}-\bar{x}_2) \quad \text{where} \quad \bar{x} = (x_L+x_H)/2. \qquad \ldots \quad (3.1.7)$$

The ratio of (3.1.6) to (3.1.7) provides some guidance on the proportional error in using simply this average difference.

If it is decided not to use the average difference, the differences $(\tau_1-\tau_2)$ for specified x can be estimated by standard methods from the separate regressions of y on x in the two samples.

To examine the relation between $(\tau_1-\tau_2)$ and x from pair-matched samples, it is natural to look at the regression of $(y_{1j}-y_{2j})$ on $\bar{x}_{.j} = (x_{1j}+x_{2j})/2$. However, from the models (3.1.1) it turns out that

$$E\{(y_{1j}-y_{2j})_m \mid \bar{x}_{.j} = x\} = (\mu_1-\mu_2)-\beta_1\eta_1+\beta_2\eta_2+(\beta_1-\beta_2)\bar{x}_{.j}+(\beta_1+\beta_2)E(d_j \mid \bar{x}_{.j} = x)$$

where $d_j = (x_{1j}-x_{2j})/2$. With $\eta_1 \neq \eta_2$ or $\sigma_1^2 \neq \sigma_2^2$, it appears that $E(d_j \mid \bar{x}_{.j} = x) \neq 0$, so that this method does not estimate the relation (3.1.2) without bias. The bias should be unimportant with tight matching, but would require Monte Carlo investigation.

3.2. *Regression non-linear*. Comparison of the performance of pair-matching with linear regression adjustment is of great interest here, since this is the situation in which, intuitively, pair-matching may be expected to be superior. Use of both weapons—linear regression on matched samples—is also relevant.

Monte Carlo comparisons were made, (Rubin, 1973b), for the monotonic non-linear functions $y = e^{\pm \frac{1}{2}x}$ and $e^{\pm x}$ and the random order nearest available matching method described earlier in Section 2.4. In such studies it is hard to convey to the reader an idea of the amount of non-linearity present. One measure will be quoted. For convenience, the Monte Carlo work was done with $\eta_1+\eta_2 = 0$ and $(\sigma_1^2+\sigma_2^2)/2 = 1$. Thus in the average population, x is $N(0, 1)$. In this population the percent of the variance of $y = e^{\pm ax}$ that is attributable to its *linear* component of regression on x is $100a^2/(e^{a^2}-1)$. For $a = \pm \frac{1}{2}, \pm 1$, respectively, 12% and 41% of the variance of y are *not* attributable to the linear component. From this viewpoint, $y = e^{\pm \frac{1}{2}x}$ might be called moderately and $y = e^{\pm x}$ markedly non-linear.

With regression adjustments on random samples, the regression coefficient used in the results presented here is $\hat{\beta}_p$, the pooled within-samples estimate. With regression adjustments on matched samples, the results are for $\hat{\beta}_d$, as would be customary in practice. Rubin (1973b) has investigated use of $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_p$ and $\hat{\beta}_d$ in both situations. He found $\hat{\beta}_p$ in the unmatched case and $\hat{\beta}_d$ in the matched case to be on the whole the best choices.

The results were found to depend markedly on the ratio σ_1^2/σ_2^2. Table 3.2.1 presents percent reductions in bias for $\sigma_1^2/\sigma_2^2 = 1$, the simplest and possibly the most common case. Linear regression on random samples performs admirably, with only a trifling over-adjustment for $y = e^{\pm x}$. Matching is inferior, particularly for $B > \frac{1}{2}$, even with a reservoir of size $4n$ from which to seek matches. Linear regression on matched samples does about as well as linear regression on random samples. Results are for $n = 50$.

Turning to the case $\sigma_1^2/\sigma_2^2 = \frac{1}{2}$ in which better matches can be obtained, note first that linear regression on random samples gives wildly erratic results which call for a rational explanation, sometimes markedly overcorrecting or even (with

$B = \frac{1}{2}$ for e^{x}) greatly increasing the original bias.[4] Matching alone does well, on the average about as well as with a linear relation (Table 3.4.1) when $\sigma_1^2/\sigma_2^2 = \frac{1}{2}$. Linear regression on matched samples is highly effective, being slightly better than matching alone.

TABLE 3.2.1. PERCENT REDUCTION IN BIAS OF $y(\sigma_1^2/\sigma_2^2 = 1)$: x NORMAL

		$B = \frac{1}{4}$				$B = \frac{1}{2}$			
method*	r	$e^{x/2}$	$e^{-x/2}$	e^{x}	e^{-x}	$e^{x/2}$	$e^{-x/2}$	e^{x}	e^{-x}
R		100	100	101	101	102	101	103	103
M	2	83	90	70	106	74	94	68	96
	3	90	101	79	104	87	98	75	100
	4	94	101	87	103	93	98	84	100
RM	2	99	103	100	106	102	100	106	101
	3	100	101	100	103	100	100	102	101
	4	100	101	100	102	100	100	101	101

		$B = \frac{3}{4}$				$B = 1$			
method	r	$e^{x/2}$	$e^{-x/2}$	e^{x}	e^{-x}	$e^{x/2}$	$e^{-x/2}$	e^{x}	e^{-x}
R		101	101	104	104	103	103	106	106
M	2	62	87	47	94	53	88	30	91
	3	81	96	68	99	70	93	55	97
	4	87	98	76	100	79	96	65	99
RM	2	103	99	110	100	104	99	113	99
	3	103	99	106	100	103	100	109	100
	4	101	100	103	100	103	100	106	99

*R denotes *linear* regression adjustment on random samples ($\hat{\beta}_p$).

M denotes 'nearest available' matching.

RM denotes *linear* regression adjustment on matched samples ($\hat{\beta}_d$).

[4]The most extreme results follow from the nature of the function $e^{\pm x}$. Consider e^{x}. Its mean value in population i is $e^{(\sigma_i^2/2+\eta_i)}$. For $B = \frac{1}{4}$, with $\eta_1 = \frac{1}{8}$, $\eta_2 = -\frac{1}{8}$, $\sigma_1^2 = \frac{2}{3}$, $\sigma_2^2 = \frac{4}{3}$, the initial bias in y is negative. Since $\eta_1 > \eta_2$ and $\hat{\beta}_p$ is positive, the regression adjustment greatly increases this negative bias, giving -304% reduction. For $B = \frac{1}{2}$, the initial bias is positive but small, so that regression greatly overcorrects, giving 282% reduction. For $B = \frac{3}{4}$, 1, the initial biases are larger and the over-correction not so extreme (170%, 139%).

With $\sigma_1^2/\sigma_2^2 = 2$ (Table 3.2.3), linear regression alone performs just as erratically as with $\sigma_1^2/\sigma_2^2 = \frac{1}{2}$, the results being in fact the same if e^{ax} is replaced by e^{-ax}. As expected from the results in Section 2.3, matching alone is poor. In most cases, regression on matched samples is satisfactory, except for failures with $e^{-x/2}$ and e^{-x} when $B = \frac{1}{4}$ or $\frac{1}{2}$.

TABLE 3.2.2. PERCENT REDUCTION IN BIAS OF y ($\sigma_1^2/\sigma_2^2 = \frac{1}{2}$, THE EASIER CASE FOR MATCHING); x NORMAL

		$B = \frac{1}{4}$				$B = \frac{1}{2}$			
method	r	$e^{x/2}$	$e^{-x/2}$	e^{x}	e^{-x}	$e^{x/2}$	$e^{-x/2}$	e^{x}	e^{-x}
R		206	62	−304	48	146	80	292	72
M	2	95	99	106	100	96	99	93	99
	3	99	100	103	100	98	100	94	100
	4	99	100	102	100	99	100	97	100
RM	2	102	100	96	100	101	100	108	100
	3	100	100	100	100	100	100	101	101
	4	100	100	100	100	100	100	100	100

		$B = \frac{3}{4}$				$B = 1$			
method	r	$e^{x/2}$	$e^{-x/2}$	e^{x}	e^{-x}	$e^{x/2}$	$e^{-x/2}$	e^{x}	e^{-x}
R		123	90	170	88	113	96	139	102
M	2	89	96	86	98	76	91	69	96
	3	97	100	94	100	94	98	90	99
	4	99	100	97	100	97	99	94	100
RM	2	103	99	113	100	106	99	118	99
	3	100	100	102	100	99	99	106	100
	4	100	101	101	100	101	100	102	100

3.3. *Regressions parallel but quadratic*. Some further insight into the performances of these methods is obtained by considering the model

$$y_{ij} = \tau_i + \beta x_{ij} + \delta x_{ij}^2 + e_{ij}. \qquad \ldots \quad (3.3.1)$$

It follows that

$$E_c(\bar{y}_1 - \bar{y}_2) = (\tau_1 - \tau_2) + \beta(\bar{x}_1 - \bar{x}_2) + \delta(\bar{x}_1^2 - \bar{x}_2^2) + \delta(s_1^2 - s_2^2) \qquad \ldots \quad (3.3.2)$$

where $s_i^2 = \Sigma\,(x_{ij} - \bar{x}_i)^2/n$. Hence the initial bias in random samples is, unconditionally,

$$(\eta_1 - \eta_2)[\beta + \delta(\eta_1 + \eta_2)] + \delta(\sigma_1^2 - \sigma_2^2) \qquad \ldots \quad (3.3.3)$$

$$= (\eta_1 - \eta_2)\beta + \delta(\sigma_1^2 - \sigma_2^2) \qquad \ldots \quad (3.3.4)$$

TABLE 3.3.3. PERCENT REDUCTION IN BIAS OF y ($\sigma^2/\sigma_1^2 = 2$, THE HARDER CASE FOR MATCHING); x NORMAL

		$B = \frac{1}{4}$				$B = \frac{1}{2}$			
method	r	$e^{x/2}$	$e^{-x/2}$	e^{x}	e^{-x}	$e^{x/2}$	$e^{-x/2}$	e^{x}	e^{-x}
R		62	206	48	−304	80	146	72	298
M	2	46	121	26	−50	45	81	30	123
	3	66	120	51	−48	60	80	43	118
	4	70	121	65	1	65	84	48	126
RM	2	90	177	90	−90	100	111	107	171
	3	98	149	98	−30	100	108	106	147
	4	96	140	94	−5	100	107	104	146

		$B = \frac{3}{4}$				$B = 1$			
method	r	$e^{x/2}$	$e^{-x/2}$	e^{x}	e^{-x}	$e^{x/2}$	$e^{-x/2}$	e^{x}	e^{-x}
R		90	123	88	170	96	113	108	130
M	2	36	72	22	90	31	67	18	63
	3	56	56	30	86	45	79	22	92
	4	66	80	42	100	50	84	29	96
RM	2	106	103	130	115	100	90	137	104
	3	108	102	111	112	106	100	119	104
	4	108	101	111	97	106	99	119	108

where without loss of generality we have assumed $\eta_1+\eta_2 = 0$.

Even though $\eta_1 > \eta_2$, if $\delta > 0$ (as appropriate for the positive exponential function) (3.3.4) shows that if $\sigma_1^2 < \sigma_2^2$, the initial bias might be small or even negative. This may indicate why some erratic results appear in the percent reduction in bias with non-linear functions.

From (3.3.2), the remaining bias in matched samples is

$$(\eta_1 - E_m(\bar{x}_2))[\beta+\delta(\eta_1+E_m(\bar{x}_2))]+\delta\{\sigma_1^2-E_m(s_2^2)\}. \qquad \ldots \quad (3.3.5)$$

The second term should be minor if the samples are relatively well matched. The first term suggests that in this case the percent reduction in bias should approximate that for parallel-linear regressions if $|\delta/\beta|$ is small. For example, let $\sigma_1^2 = \sigma_2^2 = 1$ and θ be the percent reduction in bias for y linear. From (3.3.4) and (3.3.5), the percent reduction in bias for y quadratic works out approximately as

$$-(100-\theta)\,\frac{\delta}{\beta}\,[\eta_1-E_m(\bar{x}_2)] = \theta\left[1-\frac{\delta}{\beta}\left(1-\frac{\theta}{100}\right)B\right].$$

For regression adjusted estimates on random samples, $E_c(\hat{\beta}_p)$ may be expressed as

$$E_c(\hat{\beta}_p) = \beta+\delta\left[\frac{2x_1s_1^2+2x_2s_2^2}{s_1^2+s_2^2}\right]+\frac{\delta(k_{31}+k_{32})}{s_1^2+s_2^2}$$

where $k_{3i} = \Sigma(x_{ij}-\bar{x}_i)^3/n$ is the sample third moment. From (3.3.2) it follows that the residual bias in the regression adjusted estimate on random samples is conditionally

$$= E_c[(\bar{y}_1-\bar{y}_2)-\hat{\beta}_p(x_1-x_2)]-(\tau_1-\tau_2)$$

$$= \delta(s_1^2-s_2^2)+\delta(x_1-x_2)\left[(x_1+x_2)-\frac{2(x_1s_1^2+x_2s_2^2)}{s_1^2+s_2^2}\right]-\delta(x_1-x_2)(k_{31}+k_{32})/(s_1^2+s_2^2).$$

For a symmetric or near-symmetric distribution of x in both populations the third term becomes unimportant. The first two terms give

$$\delta(s_1^2-s_2^2)[1-(x_1-x_2)^2/(s_1^2+s_2^2)].$$

The average residual bias in large random samples after regression adjustment is therefore, for x symmetric and $(\sigma_1^2+\sigma_2^2)/2 = 1$,

$$\delta(\sigma_1^2-\sigma_2^2)\left(1-\frac{(\eta_1-\eta_2)^2}{2}\right).$$

This formula suggests, as we found for $e^{\pm ax}$, that with a symmetric x and $\sigma_1^2 = \sigma_2^2$, linear regression adjustment in random samples should remove essentially all the bias when the relation between y and x can be approximated by a quadratic function. The further indication that with $\sigma_1^2 \neq \sigma_2^2$ the residual bias is smaller absolutely as $\eta_1-\eta_2$ increases towards 1 is at first sight puzzling, but consistent, for example, with the Monte Carlo results for $e^{x/2}$ and e^x when $\sigma_1^2/\sigma_2^2 = 2$ in Table 3.2.3.

To summarize for the exponential and quadratic relationships : If it appears that $\sigma_1^2 \simeq \sigma_2^2$ and x is symmetric (points that can be checked from initial data on x) linear regression adjustment on random samples removes all or nearly all the bias. Pair matching alone is inferior. Generally, regression adjustment on pair-matched samples is much the best performer, although sometimes failing in extreme cases. An explanation for this result is given in Rubin (1973b) but is not summarized here because it is quite involved. Further work on adjustment by quadratic regression, on other curvilinear relations, and on the cases $\sigma_1^2/\sigma_2^2 = \frac{3}{4}, \frac{4}{3}$ would be informative.

Before leaving the problem of non-linear regressions, we indicate how the above results can be extended to non-linear response surfaces other than quadratic. Let

$$y_{ij} = \tau_i+g(x_{ij})+e_{ij}$$

where $g(\cdot)$ is the regression surface. Since $\hat{\beta}_p$ may be written as $\sum_i\sum_j(y_{ij}-\bar{y}_i)(x_{ij}-\bar{x}_i)/\sum_i\sum_j(x_{ij}-\bar{x}_i)^2$ the limit of $\hat{\beta}_p$ in large random samples is

$$[\mathrm{cov}_1(x, g(x))+\mathrm{cov}_2(x, g(x))]/[\mathrm{var}_1(x)+\mathrm{var}_2(x)]$$

where cov_i and var_i are the covariances and variances in population i. Hence the regression adjusted estimate in large random samples has limiting residual bias

$$E_1(g(x)) - E_2(g(x)) - (\eta_1 - \eta_2)[\mathrm{cov}_1(x, g(x)) + \mathrm{cov}_2(x, g(x))]/[\mathrm{var}_1(x) + \mathrm{var}_2(x)].$$

This quantity can be calculated analytically for many distributions and regression surfaces $g(\cdot)$, (e.g., normal distributions and exponential $g(\cdot)$). In addition, if g is expanded in a Taylor series, the residual bias in random or matched samples may be expressed in terms of the moments of x in random and matched samples.

3.4. *Errors of measurement in x.* In this section we assume that y has the same linear regression on the correctly measured x (denoted by X) in both populations, but that matching or regression adjustment is made with respect to a fallible $x_{ij} = X_{ij} + u_{ij}$, where u_{ij} is an error of measurement. As in Section 3.1 the model is

$$y_{ij} = \mu_i + \beta(X_{ij} - \eta_i) + e_{ij} \tag{3.4.1}$$

and the expected bias in $(\bar{y}_1 - \bar{y}_2)$ in random samples is as before $\beta(\eta_1 - \eta_2)$.

To cover situations that arise in practice it is desirable to allow (i) u_{ij} and X_{ij} to be correlated, and (ii) u_{ij} to be a biased measurement, with $E_i(u_{ij}) = \nu_i \neq 0$. A difficulty arises at this point. Even under more restrictive assumptions (u_{ij}, X_{ij} independent in a given population and $E_i(u_{ij}) = 0$), Lindley (1947) showed that the regression of y_{ij} on x_{ij} is not linear unless the cumulant generating function of the u_{ij} is a multiple of that of the X_{ij}. Lindley's results can be extended to give corresponding conditions when the u_{ij}, X_{ij} are correlated. For simplicity we assume that these extended conditions are satisfied.

The linear regressions of y on the fallible x will be written

$$y_{ij} = \mu_i + \beta^*(x_{ij} - \eta_i - \nu_i) + e^*_{ij}$$

with $E(e^*_{ij} \mid x_{ij}) = 0$. Hence, from (3.4.1),

$$\beta^* = \mathrm{cov}(yx)/\sigma_x^2 = \frac{\beta[\sigma_X^2 + \mathrm{cov}(uX)]}{\sigma_X^2 + \sigma_u^2 + 2\,\mathrm{cov}(uX)}.$$

Unless $\mathrm{cov}(uX) < -\sigma_u^2$, we have $|\beta^*| < |\beta|$, the slope of the line being damped towards zero. The results in Section 3.2 imply that in random samples or samples matched on x a regression-adjusted estimate $\bar{y}_1 - \bar{y}_2 - \hat{\beta}^*(\bar{x}_1 - \bar{x}_2)$, where $\hat{\beta}^*$ is a least squares estimate of the regression of y on the fallible x, changes the initial bias of $\bar{y}_1 - \bar{y}_2$ by the amount

$$-\beta^*(\eta_1 - \eta_2 - \nu_1 + \nu_2).$$

Since the initial bias of $\bar{y}_1 - \bar{y}_2$ in random samples is $\beta(\eta_1 - \eta_2)$, the bias of a regression adjusted estimate is

$$(\beta - \beta^*)(\eta_1 - \eta_2) - \beta^*(\nu_1 - \nu_2).$$

The last term on the right shows that biased measurements can make an additional contribution (+or −) to the residual bias. This contribution disappears if the measurement bias is the same in both populations, $\nu_1 = \nu_2$. Under this condition the percent reduction in bias due to the regression adjustment is $100\beta^*/\beta$. With the same condition, the percent reduction in bias of $(\bar{y}_1-\bar{y}_2)$ due to matching on x is easily seen to be

$$\frac{100\beta^*}{\beta}\frac{[E_m(\bar{x}_1)-\eta_2]}{(\eta_1-\eta_2)}.$$

Thus with this simple model for errors of measurement in X, their effects on matching and adjustment are similar—namely to multiply the expected percent reduction in bias by the ratio β^*/β, usually less than 1. With u, X uncorrelated, this ratio is the quantity σ_X^2/σ_x^2 often called the reliability of the measurement x (Kendall and Buckland, 1971).

If this reliability, say $(1+a^2)^{-1}$, is known, it can be used to inflate the regression adjustment to have expectation $\beta(\eta_1-\eta_2)$, (Cochran, 1968b). Thus form the "corrected" regression adjusted estimate

$$\bar{y}_1-\bar{y}_2-(1+a^2)\hat{\beta}^*(\bar{x}_1-\bar{x}_2),$$

which is unbiased for $\tau_1-\tau_2$ under this model.

In simple examples in which Lindley's conditions are not satisfied, Cochran (1970) found the regression of y on the fallible x to be monotone but curved. A thorough investigation of the effects of errors of measurement would have to attack this case also.

3.5. *Omitted confounding variable.* One of the most common criticisms of the conclusions drawn from an observational study is that they are erroneous because the investigator failed to adjust or match for another confounding variable z_{ij} that affects y. He may have been unaware of it, or failed to measure it, or guessed that its effect would be negligible. Even under simple models, however, investigation of the effects of such a variable on the initial bias and on the performance of regression and matching leads to no crisp conclusion that either rebuts or confirms this criticism in any generality.

We assume that y_{ij} has the same linear regression on x_{ij} and z_{ij} in both populations, namely

$$y_{ij} = \mu_i+\beta(x_{ij}-\eta_i)+\gamma(z_{ij}-\nu_i)+e_{ij}. \qquad \dots \quad (3.5.1)$$

Hence, assuming x and z are the only confounding variables,

$$\tau_1-\tau_2 = E(y_{1j}-y_{2j}\,|\,x_{1j} = x_{2j},\ z_{1j} = z_{2j}) = (\mu_1-\mu_2)-\beta(\eta_1-\eta_2)-\gamma(\nu_1-\nu_2)$$

and the initial bias in $(\bar{y}_1-\bar{y}_2)$ from random samples is now

$$\beta(\eta_1-\eta_2)+\gamma(\nu_1-\nu_2). \qquad \dots \quad (3.5.2)$$

Similarly, the bias in $(\bar{y}_1-\bar{y}_2)$ from samples matched on x is

$$\beta(\eta_1-E_m(\bar{x}_2))+\gamma(\nu_1-E_m(\bar{z}_2)).$$

Thus, depending on the signs of the parameters involved, the presence of z_{ij} in the model may either increase or decrease (perhaps to an unimportant amount) the previous initial bias $\beta(\eta_1-\eta_2)$. Also, even if $|\eta_1-\eta_2| > |\eta_1-\bar{E}_m(\bar{x}_2)|$ and $|\nu_1-\nu_2| > |\nu_1-\bar{E}_m(\bar{z}_2)|$, the bias of $(\bar{y}_1-\bar{y}_2)$ may be greater in matched than random samples.

Suppose now that z_{ij} has linear and parallel regressions on x_{ij} in the two populations :

$$z_{ij} = \nu_i+\lambda(x_{ij}-\eta_i)+\varepsilon_{ij}. \qquad \ldots \quad (3.5.3)$$

Then (3.5.1) may be written

$$y_{ij} = \mu_i+(\beta+\gamma\lambda)(x_{ij}-\eta_i)+\varepsilon_{ij}+e_{ij}. \qquad \ldots \quad (3.5.4)$$

In (3.5.4) we have returned to the model in Section 2.2 —same linear regression of y on x in both populations. From Section 2.1, the expected change in bias of $(\bar{y}_1-\bar{y}_2)$ due to regression adjustment on x in random samples or samples matched on x is therefore

$$-(\beta+\gamma\lambda)(\eta_1-\eta_2) \qquad \ldots \quad (3.5.5)$$

while that due to matching on x is

$$-(\beta+\gamma\lambda)[\bar{E}_m(\bar{x}_2)-\eta_2]. \qquad \ldots \quad (3.5.6)$$

As regards regression, (3.5.2) and (3.5.6) lead to the residual bias

$$\gamma[(\nu_1-\nu_2)-\lambda(\eta_1-\eta_2)]. \qquad \ldots \quad (3.5.7)$$

Thus, adjustment on x alone removes the part of the original bias coming from z that is attributable to the linear regression of z on x. If z has *identical* linear regressions on x in both populations, so that $(\nu_1-\lambda\eta_1) = (\nu_2-\lambda\eta_2)$, the residual bias is zero as would be expected. With matching in this situation, the residual bias is

$$(\beta+\gamma\lambda)[\eta_1-\bar{E}_m(\bar{x}_2)]$$

matching being less effective than regression.

With regressions of z on x parallel but not identical, the final bias with either regression or matching could be numerically larger than the initial bias, and no simple statement about the relative merits of regression and matching holds under this model.

If the regressions of z_{ij} on x_{ij} are parallel but non-linear, investigation shows that in large samples, regression and matching remove the part of the bias due to z that is attributable to the linear component of the regression of z on x.

4. Matching and adjustment by subclassification

4.0. *The two methods.* When the x-variable is qualitative, e.g. sex (M, F), it is natural to regard any male from population 1 as a match for any male from population 2 with respect to x, or more generally, any two members who fall in the same qualitative class as a match. This method is also used frequently when x is continuous, e.g. age. We first divide the range of ages that are of interest into, say, specified 5-year classes 40-44, 45-49, etc. and regard any two persons in the same age class as a match.

In matching to the sample from population 1, let n_{1j} be the number in sample 1 who fall in the j-th subclass. From the reservoir from population 2, we seek the same number $n_{2j} = n_{1j}$ in the j-th class. The average matched-pair difference, $\Sigma n_{1j}(\bar{y}_{1j}-\bar{y}_{2j})/n$ is of course the difference $(\bar{y}_1-\bar{y}_2)$ between the two matched sample means, this method being self-weighting.

With random samples from the two populations, the alternative method of adjustment by subclassification starts by classifying both samples into the respective classes. The numbers n_{1j}, n_{2j} will now usually differ. However, any weighted mean $\Sigma w_j(\bar{y}_{1j}-\bar{y}_{2j})$, with $\Sigma w_j = 1$, will be subject only to the residual within-class biases insofar as this x is concerned. In practice, different choices of the weights w_j have been used, e.g. sometimes weights directed at minimizing the variance of the weighted difference. For comparison with matching we assume the weights $w_j = n_{1j}/n$.

4.1. *Performance of the two methods*. If sample 1 and reservoir 2 or sample 2 are random samples from their respective populations, as we have been assuming throughout, the n_{1j}, n_{2j} who turn up in the final sample are a random sample from those in their population who fall in class j under either method-matching or adjustment. Consequently, with the same weights n_{1j}/n, the two methods have the same expected residual bias. (An exception is the occasional case of adjustment from initial random samples of equal sizes $n_1 = n_2 = n$, where we find $n_{2j} = 0$ in one or more subclasses, so that subclasses have to be combined to some extent for application of the 'adjustment by subclassification' method.)

With certain genuinely qualitative classifications it may be reasonable to assume that any two members of the same subclass are identical as regards the effect of this x on y. In this event, both matching and adjustment remove all the bias due to x, there being no within-class bias. But many qualitative variables like socio-economic status, degree of aggressiveness (mild, moderate, severe), represent an ordered classification of an underlying continuous variable x which at present we are unable to measure accurately. Two members of the same subclass do not have identical values of x in this event. For such cases, and for a variable like age, we assume the model

$$y_{ij} = \tau_i + u(x_{ij}) + e_{ij}, \quad i = 1, 2, \; j = 1, 2, \ldots, c, \qquad \ldots \quad (4.1.1)$$

the regression of y on x being the same in both populations, with $\tau_1-\tau_2$ not depending on the value of x.

From (4.1.1) the percent reduction in the bias of y due to adjustment by subclassification of u equals the percent reduction in the bias of u. If $u(x) = x$, this also equals the percent reduction in the bias of x. If $u(x)$ is a monotone function of x, a division of x into classes at the quantiles of x will also be a division of u into classes at the same quantiles of u. The percent reductions in bias of u and x will not, however, be equal, since these depend both on the division points and on the frequency distributions, which will differ for u and x. The approach adopted by Cochran (1968a) was to start with the case $u(x) = x$, with x normal, and then consider some non-normal distributions of x to throw some light on the situation with $u(x)$ monotone.

In subclassification, the range of x is divided into c classes at division points $x_0, x_1, \ldots, x_c$. Let $f_i(x)$ be the p.d.f.'s of x in the two populations. The overall means of x are

$$\eta_i = \int x f_i(x)\, dx$$

while in the j-th subclass the means are

$$\eta_{ij} = \int_{x_{j-1}}^{x_j} x f_i(x) dx / P_{ij}, \quad \text{where} \quad P_{ij} = \int_{x_{j-1}}^{x_j} f_i(x)\, dx.$$

The initial expected bias in x is $(\eta_1 - \eta_2)$. After matching or adjustment, the weighted mean difference in the two samples is

$$\sum_{j=1}^{c} \frac{n_{1j}}{n} (\bar{x}_{1j} - \bar{x}_{2j}). \qquad \ldots \quad (4.1.2)$$

Its average value, the expected residual bias, is

$$\sum_{j=1}^{c} P_{1j}(\eta_{1j} - \eta_{2j}). \qquad \ldots \quad (4.1.3)$$

This expression may be used in calculating the expected percent reduction in bias.

If $f_1(x)$, $f_2(x)$ differ only with respect to a single parameter it is convenient to give it the values 0 and Θ in populations 1 and 2, respectively. Expression (4.1.3) may be rewritten as

$$\sum_{j=1}^{c} P_j(0)\{\eta_j(0) - \eta_j(\Theta)\}. \qquad \ldots \quad (4.1.4)$$

A first-term Taylor expansion about 0, assuming Θ small, seems to work well for biases of practical size, (Cochran, 1968a) and leads to a useful result obtained in a related problem. From (4.1.4) the expected *residual* bias is approximately, expanding about $\Theta = 0$,

$$-\Theta \sum_{j=1}^{c} P_j(0) \frac{d\eta_j(\Theta)}{d\Theta} \qquad \ldots \quad (4.1.5)$$

the derivative being measured at $\Theta = 0$. On the other hand, the expected *initial* bias is

$$\sum_{j=1}^{c} [P_j(0)\eta_j(0) - P_j(\Theta)\eta_j(\Theta)] \simeq -\Theta \sum_{j=1}^{c} \left[P_j(0) \frac{d\eta_j(\Theta)}{d\Theta} + \eta_j(0) \frac{dP_j(\Theta)}{d\Theta}\right]. \quad \ldots \quad (4.1.6)$$

On subtracting (4.1.5) from (4.1.6), the expected proportional reduction in bias is approximately

$$\sum_{j=1}^{c} \eta_j(0) \frac{dP_j(\Theta)}{d\Theta} \Big/ \frac{d\eta(\Theta)}{d\Theta} \qquad \ldots \quad (4.1.7)$$

measured at $\Theta = 0$, where $\eta(\Theta) = \eta_2 - \sum_{j=1}^{c} P_j(\Theta)\eta_j(\Theta)$.

In particular, if $f_1(x) = f(x)$, $f_2(x) = f(x-\Theta)$, the two distributions differing only in their means, we have $\frac{d\eta}{d\Theta} = 1$ and

$$P_j(\Theta) = \int_{x_{j-1}}^{x_j} f(x-\Theta)\, dx = \int_{x_{j-1}-\Theta}^{x_j-\Theta} f(x)\, dx$$

with

$$\frac{dP_j(\Theta)}{d\Theta} = f(x_{j-1}) - f(x_j)$$

at $\Theta = 0$. From (4.1.7), the proportional reduction in bias becomes

$$\sum_{j=1}^{c} \eta_j(0)[f(x_{j-1}) - f(x_j)]. \qquad \ldots \quad (4.1.8)$$

If $f(x)$ is the unit normal distribution, (4.1.8) gives

$$\sum_{j=1}^{c} [f(x_{j-1}) - f(x_j)]^2 / P_j(0) \qquad \ldots \quad (4.1.9)$$

for the proportional reduction in bias. Expression (4.1.9) has been studied in other problems by J. Ogawa (1951) and by D. R. Cox (1957). Cox showed that it is 1 minus the ratio of the average within-class variance to the original variance of x when x is normal. For our purpose, their calculations provide (i) the optimum choices of the P_{1j}, (ii) the resulting maximum percent reductions in bias, and (iii) the percent reductions in bias with equal-sized classes $P_{1j} = 1/c$. For $c = 2-10$, the maximum percent reductions are at most about 2% higher than those for equal P_{1j}, shown in Table 4.2.1.

TABLE 4.2.1. PERCENT REDUCTIONS IN BIAS WITH EQUAL-SIZED CLASS IN POPULATION 1, x NORMAL

no. of subclasses	2	3	4	5	6	8	10
% reduction	64%	79%	86%	90%	92%	94%	96%

Calculations (Cochran, 1968a) of the percent reductions when x follows χ^2 distributions, t distributions and Beta distributions suggest that the above figures can be used as a rough guide to what to expect in practice when the classification represents an underlying continuous x. To remove 80%, 90% and 95% of the initial bias, evidently 3, 5, and 10 classes are required by this method.

5. Simple multivariate generalizations

5.1. *Parallel linear regressions.* We now consider the case of many x-variables, say $(x^{(1)}, x^{(2)}, \ldots, x^{(p)})$. Many of the previous results for one x variable have obvious analogues for p x-variables, with the p-vectors $\boldsymbol{\eta}_i$, $\boldsymbol{\beta}_i$ and $\boldsymbol{x}_{ij}$ replacing the scalars η_i, β_i and x_{ij}. However except in the cases where the adjustment removes all the bias, the conclusions are even less sharp than in the univariate case.

The simplest multivariate case occurs when y has parallel linear regressions on $\boldsymbol{x}$ in both populations

$$y_{ij} = \mu_i + \boldsymbol{\beta}(\boldsymbol{x}_{ij} - \boldsymbol{\eta}_i)' + e_{ij}. \qquad \ldots \quad (5.1.1)$$

The regressions of y on x in the two populations are parallel "planes" with a constant difference of height

$$E(y_{1j} - y_{2j} \mid \boldsymbol{x}_{1j} = \boldsymbol{x}_{2j}) = (\mu_1 - \mu_2) - \boldsymbol{\beta}(\boldsymbol{\eta}_1 - \boldsymbol{\eta}_2)'. \qquad \ldots \quad (5.1.2)$$

If $(x^{(1)}, \ldots, x^{(p)})$ are the only confounding variables, this constant difference is the treatment difference, $\tau_1 - \tau_2$. From (5.1.1) it follows that conditionally on the values of the $\boldsymbol{x}_{ij}$ in two samples, chosen either randomly or only on the basis of the x-variables,

$$E_c(\bar{y}_1 - \bar{y}_2) = \tau_1 - \tau_2 + \boldsymbol{\beta}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'.$$

The expected bias of $\bar{y}_1 - \bar{y}_2$ in random samples is

$$E_r(\bar{y}_1 - \bar{y}_2) - (\tau_1 - \tau_2) = \boldsymbol{\beta}(\boldsymbol{\eta}_1 - \boldsymbol{\eta}_2)'. \qquad \ldots \quad (5.1.3)$$

Notice that since $\boldsymbol{\beta}$ and $(\boldsymbol{\eta}_1 - \boldsymbol{\eta}_2)$ are vectors the initial bias in $(\bar{y}_1 - \bar{y}_2)$ may be zero even if $\boldsymbol{\beta} \neq 0$ and $(\boldsymbol{\eta}_1 - \boldsymbol{\eta}_2) \neq 0$.

In random P_1 and matched P_2 samples, the bias is

$$E_m(\bar{y}_1 - \bar{y}_2) - (\tau_1 - \tau_2) = \boldsymbol{\beta}(\boldsymbol{\eta}_1 - E_m(\bar{\boldsymbol{x}}_2))'. \qquad \ldots \quad (5.1.4)$$

Formally, the percent reduction in bias is the natural extension of the univariate result,

$$100\boldsymbol{\beta}(E_m(\bar{\boldsymbol{x}}_2) - \boldsymbol{\eta}_2)' / \boldsymbol{\beta}(\boldsymbol{\eta}_1 - \boldsymbol{\eta}_2)'. \qquad \ldots \quad (5.1.5)$$

But the bias in matched samples may be greater than in random samples even with $E_m(\bar{\boldsymbol{x}}_2)$ closer to $\boldsymbol{\eta}_1$ in all components than $\boldsymbol{\eta}_2$ is to $\boldsymbol{\eta}_1$(e.g. $(\boldsymbol{\eta}_1 - \boldsymbol{\eta}_2) = (1, -1)$, $(\boldsymbol{\eta}_1 - E_m(\bar{\boldsymbol{x}}_2)) = (\frac{1}{2}, \frac{1}{2})$, $\boldsymbol{\beta} = (1, 1)$), which give initial bias 0 and matched sample bias 1.

The regression adjusted estimate is

$$\hat{\tau}_1 - \hat{\tau}_2 = (\bar{y}_1 - \bar{y}_2) - \hat{\boldsymbol{\beta}}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)' \qquad \ldots \quad (5.1.6)$$

where $\hat{\boldsymbol{\beta}}$ is the vector of estimated regression coefficients of y on $\boldsymbol{x}$. Under this model, $E_c(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, for $\hat{\boldsymbol{\beta}}_p$, $\hat{\boldsymbol{\beta}}_1$, $\hat{\boldsymbol{\beta}}_2$. Thus, for any of these $\hat{\boldsymbol{\beta}}$ and samples either random or matched on x, (5.1.1) and (5.1.6) show that the regression adjusted estimate is unbiased :

$$E_c(\hat{\tau}_1 - \hat{\tau}_2) = \mu_1 - \mu_2 - \boldsymbol{\beta}(\boldsymbol{\eta}_1 - \boldsymbol{\eta}_2)' = \tau_1 - \tau_2.$$

5.2. *Non-parallel linear regressions*. As in the univariate case, the regressions of y on x may not be parallel. Assume the objective is to estimate $(\tau_1-\tau_2)$ averaged over some standard population with mean x vector η_s (e.g. $\eta_s = \eta_1$ if P_1 is considered the standard). From the multivariate version of (3.1.2), assuming x are the only confounding variables we have

$$E_s(\tau_1-\tau_2) = \mu_1-\mu_2+\beta_1(\eta_s-\eta_1)'-\beta_2(\eta_s-\eta_2)'. \quad \dots \quad (5.2.1)$$

In random samples $\bar{y}_1-\bar{y}_2$ has expectation $\mu_1-\mu_2$ and thus the initial bias is

$$-\beta_1(\eta_s-\eta_1)'+\beta_2(\eta_s-\eta_2)'.$$

If $\eta_s = \eta_1$, this initial bias becomes $\beta_2(\eta_1-\eta_2)'$.

For random samples or samples selected solely on x

$$\begin{aligned} E_c(\bar{y}_1-\bar{y}_2) &= \mu_1-\mu_2+\beta_1(\bar{x}_1-\eta_1)'-\beta_2(\bar{x}_2-\eta_2)' \\ &= E_s(\tau_1-\tau_2)+\beta_1(\bar{x}_1-\eta_s)'-\beta_2(\bar{x}_2-\eta_s)'. \quad \dots \quad (5.2.2) \end{aligned}$$

If $\eta_s = \eta_1$, and sample 1 is a random sample, the bias of $\bar{y}_1-\bar{y}_2$ is $\beta_2(\eta_1-E_m(\bar{x}_2))'$ while the initial bias is $\beta_2(\eta_1-\eta_2)'$. By comparison with (5.1.3) and (5.1.4) it follows that when population 1 is chosen as the standard the effect of matching on bias reduction is the same whether the regressions are parallel or not.

Now consider the regression estimate. Since (5.2.2) gives the conditional bias of $\bar{y}_1-\bar{y}_2$ it would seem reasonable to estimate this bias using the usual within-sample least squares estimates of β_1 and β_2 (and an estimate of η_s if necessary) and forming the regression adjusted estimate

$$\bar{y}_1-\bar{y}_2-\hat{\beta}_1(\bar{x}_1-\hat{\eta}_s)'+\hat{\beta}_2(\bar{x}_2-\hat{\eta}_s)' \quad \dots \quad (5.2.3)$$

which is an unbiased estimate of $E_s(\tau_1-\tau_2)$ under the linear regression model. If $\eta_s = \eta_1$ and the first sample is random, this estimate is the natural extension of the univariate result,

$$\bar{y}_1-\bar{y}_2-\hat{\beta}_2(\bar{x}_1-\bar{x}_2)'.$$

If a single summary of the effect of the treatment is not adequate, one could examine the estimated effect at various values of x using (5.2.3) where η_s is replaced by the values of x of interest.

5.3. *Non-linear regressions*. If y has non-linear parallel regressions on x, expressed by the function $g(x)$, the initial bias, $E_1(g(x))-E_2(g(x))$, depends on the higher moments of the distributions of x in P_1 and P_2 (e.g. the covariance matrices Σ_1 and Σ_2 if x is normal) as well as the means. The large sample limit of the pooled regression adjusted estimate in random samples is

$$E_1(g(x))-E_2(g(x))-(\eta_1-\eta_2)[\Sigma_1+\Sigma_2]^{-1}C''$$

where the k-th component of the p-vector C is $\text{cov}_1(x^{(k)}g(x))+\text{cov}_2(x^{(k)}g(x))$.

This quantity, as well as similar quantities for the case of parallel, non-linear regressions, can be obtained analytically for many distributions and regression functions. As far as we know, no work has been done on this problem or the more difficult one involving matched samples, in which case the distribution of x in matched samples may not be analytically tractable. Expanding $g(x)$ in a Taylor series would enable one to expand the limiting residual bias in terms of the moments of x in random and matched samples (matched moments for the regression adjusted estimate based on matched pairs).

5.4. *Errors of measurement in x*. Assume that y has parallel linear regressions on the correctly measured matching variables X, that X are the only confounding variables, but that matching and regression adjustment are done on the fallible $x = X+u$. Hence

$$y_{ij} = \mu_i + \beta(X_{ij} - \eta_i)' + e_{ij}$$

and the initial bias in random samples is $\beta(\eta_1 - \eta_2)'$. If y has a linear regression on the fallible x (e.g. y, X, u are multivariate normal), let

$$y_{ij} = \mu_i + \beta^*(x_{ij} - \eta_i - \nu_i) + e^*_{ij}$$

where $E(e^*_{ij} \mid x_{ij}) = 0$, $E(u_{ij}) = \nu_i$, and $\beta^* = \Sigma_x^{-1} \operatorname{cov}(y, x)$, where Σ_x is the covariance matrix of the x variables.

A regression adjusted estimate based on random samples or samples matched on x changes the initial bias by the amount $-\beta^*(\eta_1 - \eta_2 - \nu_1 + \nu_2)'$, and thus the bias of a regression adjusted estimate is

$$(\beta - \beta^*)(\eta_1 - \eta_2)' - \beta^*(\nu_1 - \nu_2)'. \qquad \dots \quad (5.4.1)$$

Some simple results can be obtained for the special case when $\nu_1 = \nu_2$ (equally biased measurements in both populations), X and u are uncorrelated, and the covariance matrix of u is proportional to the covariance matrix of X, say $a^2\Sigma_X$. With the latter two conditions β^* becomes $(1+a^2)^{-1}\beta$. This result and $\nu_1 = \nu_2$ imply that (5.4.1) becomes

$$\frac{a^2}{1+a^2}\, \beta(\eta_1 - \eta_2)'$$

and the percent reduction in bias due to regression adjustment is

$$100/(1+a^2),$$

as in the univariate case, since $1/(1+a^2)$ corresponds to the reliability, which we are assuming to be uniform for all variables. Under this same set of special conditions, the percent reduction in bias due to matching on x would be

$$\frac{100}{1+a^2}\, \beta(E_m(x_2) - \eta_2)'/\beta(\eta_1 - \eta_2)'.$$

Under models different from the above special case, clearcut results appear more difficult to obtain, and the percent reduction in bias for regression adjustment or matching is not necessarily between 0 and 100 percent even with $\mathbf{v}_1 = \mathbf{v}_2$, $\boldsymbol{X}$ and $\boldsymbol{u}$ uncorrelated, and all u_i independent.

If one knew $\boldsymbol{\Sigma}_u$, one could form a "corrected" regression-adjusted estimate that is in large samples unbiased for $\tau_1-\tau_2$. That is, assuming $\boldsymbol{x}$ and $\boldsymbol{u}$ are uncorrelated, form

$$\bar{y}_1-\bar{y}_2-\hat{\boldsymbol{\Sigma}}_X^{-1}\hat{\boldsymbol{\Sigma}}_x\hat{\boldsymbol{\beta}}^*(\bar{\boldsymbol{x}}_1-\bar{\boldsymbol{x}}_2)' \qquad \ldots \quad (5.4.2)$$

where $\hat{\boldsymbol{\beta}}^*$ is the usual least squares estimate of the regression of y on $\boldsymbol{x}$, $\hat{\boldsymbol{\Sigma}}_x$ is the estimated within group covariance matrix of $\boldsymbol{x}$ and $\hat{\boldsymbol{\Sigma}}_X = \hat{\boldsymbol{\Sigma}}_x-\boldsymbol{\Sigma}_u$. In the special case when $\boldsymbol{\Sigma}_u = a^2\boldsymbol{\Sigma}_X$, the estimate simplifies to the analogue of the univariate result if a^2 is known

$$\bar{y}_1-\bar{y}_2-(1+a^2)\hat{\boldsymbol{\beta}}^*$$

which is unbiased for $\tau_1-\tau_2$.

5.5. *Omitted confounding variables*. Assume that y has parallel regressions on $(\boldsymbol{x}, \boldsymbol{z})$ in the populations but that matching and/or adjustment is done on the $\boldsymbol{x}$ variables alone. Also assume that $\boldsymbol{x}$ and $\boldsymbol{z}$ are the only confounding variables. This multivariate case is very similar to the univariate one of Section 3.5 and the multivariate analogs of all the formulas follow in an obvious manner. The basic result is that if $\boldsymbol{z}$ has a linear regression on $\boldsymbol{x}$, $\boldsymbol{z}$ can be decomposed into $\boldsymbol{z}_a$ along $\boldsymbol{x}$ and $\boldsymbol{z}_0$ orthogonal to $\boldsymbol{x}$, and adjustment on $\boldsymbol{x}$ is also adjustment on $\boldsymbol{z}_a$ but does not affect $\boldsymbol{z}_0$.

6. Some multivariate generalizations of univariate matching methods

6.1. *Caliper matching*. Thus far we have not discussed any specific multivariate matching methods. The obvious extension of caliper matching is to seek in reservoir 2 a match for each x_{1j} such that $|x_{1j}^{(k)}-x_{2j}^{(k)}| < c_k$ for $k = 1, 2, \ldots, p$. This method is used in practice, the difficulty being the large size of reservoir needed to find matches.

The effect of this method on $E_m(\bar{y}_1-\bar{y}_2)$ could be calculated from univariate results if all x were independently distributed in P_2 (this restriction will be relaxed shortly). This follows because selection on x_{ij} from P_2 would not affect the other x variables, and so the percent reduction in the bias of the variate $x^{(k)}$ under this method would be the same as that under the univariate caliper matching $|x_{1j}^{(k)}-x_{2j}^{(k)}| < c_k$. From these p percent reductions, the percent reduction in bias could be calculated for any y that is linear in the x. For example, with $p = 2$, let $B = 0.5$ and $\sigma_1^2/\sigma_2^2 = \frac{1}{2}$ for $x^{(1)}$, while $B = 0.25$ and $\sigma_1^2/\sigma_2^2 = 2$ for $x^{(2)}$. Then if $c_1 = 0.4\sqrt{(\sigma_1^2+\sigma_2^2)/2}$ and $c_2 = 0.8\sqrt{(\sigma_1^2+\sigma_2^2)/2}$, the reductions for $x^{(1)}$ and $x^{(2)}$ from Table 2.3.1 are about 96% and 77%. That for $x^{(1)}+x^{(2)}$, for instance, is about $[(.96)(.5)+(.77)(.25)]/(.75) = 90\%$.

With this approach, an attempt to select the c_k from initial estimates of $\eta_1^{(k)}-\eta_2^{(k)}$ and σ_1^2/σ_2^2 for $x^{(k)}$ so that matching gives the same percent reduction in bias for each $x^{(k)}$ has some appeal, particularly when matching for more than one y or when it is uncertain which $x^{(k)}$ are more important. Whenever this property does not hold, matching can *increase* the bias for some y's linear in x. For instance, in the preceding example matching would increase the bias for $x^{(1)}-2x^{(2)}$, whose bias is initially zero.

In general of course, the matching variables are not independently distributed in P_2, but if they are normally distributed (or more generally spherically distributed, Dempster, 1969) there exists a simple linear transformation

$$z = xH \quad \text{where} \quad H'H = \Sigma_2^{-1} \qquad \ldots \quad (6.1.1)$$

such that the z are independently distributed in P_2. Hence, assuming (1) x normal in P_2, (2) a large sample from P_2 so that H is essentially known and all matches can be obtained, and (3) the caliper matching method defined above is used on the $z = xH$ variables, Table 2.3.1 can be used to calculate the percent reduction in bias for each of the $z^{(k)}$. Also, from these p percent reductions in bias, the percent reduction in bias can be calculated for any linear combination of the $z^{(k)}$, such as any $x^{(k)}$ or any y that is linear in x.

We consider caliper matching on the transformed variables to be a reasonable generalization of univariate caliper matching to use in practice. Caliper matching on the original x variables defines a fixed p-dimensional "rectangular" neighborhood about each x_{1j} in which an acceptable match can be found. If caliper matching is used on the z-variables, a neighborhood is defined about each x_{1j} that in general is no longer a simple rectangle with sides perpendicular to the x-variables but a p-dimensional parallelopiped whose sides are not perpendicular to the x-variables but to the p linear combinations of the x-variables corresponding to the z. Since the original choice of a rectangular neighborhood (e.g. rather than a circular one) was merely for convenience, the neighborhood defined by the z calipers should be just as satisfactory.

6.2. *Categorical matching*. As a second example of a commonly used matching method for which we can apply the univariate results, assume the categorical matching method of Section 4 is used with c_k categories for each matching variable, the final match for each member of the first sample being chosen from the members of the second sample lying in the same categories on all variables. If this matching is performed on the transformed variables z given in (6.1.1), normality is assumed, and the reservoir is large, Table 4.2.1 can be used to calculate the percent reduction in bias of each $z^{(k)}$ in the final matched sample, and thus of each $x^{(k)}$ or any y linear in x. Actually Table 4.2.1 requires the ratio of variances to be 1 and B moderate or small but could be extended to include more cases.

By adjusting the number of categories used per matching variable $z^{(k)}$ as a function of $E_1(z^{(k)})-E_2(z^{(k)})$ and $\mathrm{var}_1(z^{(k)})/\mathrm{var}_2(z^{(k)})$ one can obtain approximately the same percent reduction in bias of any y that is linear in x.

6.3. *Discriminant matching.* As a final example of multivariate matching methods for which some of the previous univariate results are applicable, assume the transformation in (6.1.1) will be used with H defined so that $(\bar{x}_1-\bar{x}_2)H \propto (1, 0, \ldots, 0)$. Univariate matches are then obtained on $z^{(1)}$, the best linear discriminant with respect to the Σ_2 inner product, as suggested by Rubin (1970). Note with this method there is no (mean) bias orthogonal to the discriminant (i.e. $E_1 z^{(k)} = E_2 z^{(k)}$, $k = 2, \ldots, p$); hence, if the x are normal in P_2 (so that $z^{(1)}$ and $(z^{(2)}, \ldots, z^{(p)})$ are independent), the percent reduction in bias for any linear function of the x equals the percent reduction in bias of $z^{(1)}$.

Tables 2.3.1, 2.5.1, 2.4.1, or 4.2.1 can then be used to calculate the percent reduction in bias for each $z^{(k)}$ when univariate caliper, mean, nearest available or categorical matching is used on the discriminant. In using these tables σ_1^2/σ_2^2 is the ratio of the $z^{(1)}$ variances in P_1 and P_2, $(\bar{x}_1-\bar{x}_2)\Sigma_2^{-1}\Sigma_1\Sigma_2^{-1}(\bar{x}_1-\bar{x}_2)'/(\bar{x}_1-\bar{x}_2)\Sigma_2^{-1}(\bar{x}_1-\bar{x}_2)'$, and B is the number of standard deviations between the means of $z^{(1)}$ in P_1 and P_2, $(\bar{x}_1-\bar{x}_2)\Sigma_2^{-1}(\bar{x}_1-\bar{x}_2)'/\sqrt{\frac{1}{2}(\sigma_1^2+\sigma_2^2)}$. Note that for many matching variables, this B could be quite large even if the means of each matching variable are moderately similar in P_1 and P_2.

Discriminant matching has several appealing properties :

(1) it is easy to control the sizes of the final matched samples to be exactly of size n;

(2) if x is approximately normal in P_2 the method should do a good job of reducing bias of any y linear in x, even for a modest reservoir; this follows from an examination of Tables 2.4.1 and 2.5.1 ;

(3) if x is approximately normal in both P_1 and P_2 with $\Sigma_1 \simeq \Sigma_2$, pair matching should do a good job of reducing the bias of any type of regression when the reservoir is large and/or when combined with regression adjustment.

The third point follows from the fact that if x is normal in P_1 and P_2 with $\Sigma_1 = \Sigma_2$, orthogonal to the discriminant the distributions of the matching variables are identical in P_1 and P_2 and unaffected by the matching. Hence, for any y, all bias is due to the different distributions of the discriminant, and Tables 3.2.1-3.2.3 indicate that with moderate r, matching and regression adjustment remove much of this bias; also when $r \to \infty$ the distributions of all matching variables will be the same in the matched samples if nearest available matching is used and x is normal with $\Sigma_1 = \Sigma_2$.

In addition, if one had to choose one linear combination of the x along which a non-linear y is changing most rapidly, and thus on which to obtain close pair matches, the discriminant seems reasonable since the matching variables were presumably chosen not only because their distributions differ in P_1 and P_2 but also because they are correlated with y.

Of course, the joint distributions of matching variables are not assured to be similar in the matched samples, as they would be with pair matches having tight calipers or with a large number of categories using the methods of Sections 6.1 or 6.2. However, the ability to find tight pair matches on all matching variables in a highly multivariate situation seems dubious even with moderately large r. The implications of these points require study.

In practice the discriminant is never known exactly. However, symmetry arguments (Rubin, 1973c) show that under normality in P_2, matching on the sample-based discriminant still yields the same percent reduction in expected bias for each $x^{(k)}$.

6.4. *Other matching methods*. There are two kinds of problems with the preceding matching methods. First, for those utilising all the x it is difficult to control the size of the final sample of matches. Thus with the caliper or categorical methods little is known about the actual reservoir size needed to be confident of obtaining a match for each member of the first sample, although an argument suggests that the ratio of reservoir to sample size for p variables i.i.d. in P_1 and P_2 is roughly the p-th power of the ratio for one variable. The use of caliper matching to obtain matched samples in a practical problem is described in Althauser and Rubin (1970).

When using mean and nearest available matching on the discriminant it is easy to control the final matched sample to have size n. However using discriminant matching, individual matched pairs are not close on all variables and they rely on specific distributional assumptions to insure that the samples are well-matched, even as $r \to \infty$.

An alternative is to try to define matching methods more analogous to the univariate nearest available matching method using some definition of "distance" between x_{1j} and x_{2j}. We might choose the n matches by ordering the x_{1j} in some way (e.g. randomly) and then assigning as a match the nearest x_{2j} as defined by some multivariate distance measure. Such methods will be called nearest available metric matching methods.

A simple class of metrics is defined by an inner product matrix, D, so that the distance from x_{1j} to x_{2j} is $(x_{1j}-x_{2j})\, D(x_{1j}-x_{2j})'$. Rather obvious choices for D are Σ_1^{-1} or Σ_2^{-1} yielding the Mahalanobis (1927) distance between x_{1j} and x_{2j} with respect to either inner product. If $\Sigma_1 \propto \Sigma_2$ and x is spherical, symmetry implies that either Mahalanobis distance yields the same percent reduction in bias for each $x^{(k)}$.

More generally, unpublished symmetry arguments (Rubin, 1973c) show that for x spherical and an inner product metric, the same percent reduction in bias is obtained for each $x^{(k)}$ if and only if

(1) The P_i covariance matrices of x orthogonal to the discriminants are proportional :

$$\Sigma_{\perp} = \Sigma_1 - \frac{1}{s_1^2}(\eta_1-\eta_2)'(\eta_1-\eta_2) = c\left[\Sigma_2 - \frac{1}{s_2^2}(\eta_1-\eta_2)'(\eta_1-\eta_2)\right]$$

where s_i^2 = the variance of discriminant in $P_i = (\eta_1-\eta_2)\Sigma_i^{-1}(\eta_1-\eta_2)'$. (Note that this implies the discriminants with respect to the P_1 and P_2 inner products are proportional).

(2) The inner product matrix D used for matching is proportional to $[\Sigma_+ + k(\eta_1-\eta_2)'(\eta_1-\eta_2)]^{-1}$ with $k \geqslant 0$ (if $k = 0$ or ∞, the inverse is a generalized inverse, Rao, 1973).

The choice of $k = \infty$ yields matching along the discriminant, $k = 0$ yields matching in the space orthogonal to the discriminant, $k = s_1^{-2}$ yields matching using the P_1 Mahalanobis distance and $k = cs_2^{-2}$ yields matching using the P_2 Mahalanobis distance. Symmetry arguments also show that under normality and condition (1), using the sample estimates of Σ_+ and $(\eta_1-\eta_2)$ gives the same percent reduction in bias for each $x^{(k)}$.

There are of course other ways to define distance between x_{1j} and x_{2j}, for example by the Minkowski metric

$$\left[\prod_{k=1}^{p} |x_{1j}^{(k)}-x_{2j}^{(k)}|^{\gamma}\right]^{1/\gamma} \quad \text{for some } \gamma > 0.$$

Nothing seems to be known about the performance of such matching methods.

A final class of methods that has not been explored might be described as sample metric matching. The simplest example would be to minimize distance between the means $\bar{x}_1$ and $\bar{x}_2$ with respect to a metric. More interesting and robust against non-linearity would be to minimize a measure of the difference between the empirical distribution functions.

7.1. *Summary comments.* This review of methods of controlling bias in observational studies has concentrated on the performance of linear regression adjustments and various matching methods in reducing the initial bias of y due to differences in the distribution of confounding variables, x, in two populations; this seemed to us the most important aspect in observational studies. We have not considered the effects of these techniques on increasing precision, as becomes the focus of interest in randomized experiments.

If the x variables are the only confounding variables, linear regression adjustment on random samples removes all the initial bias when the (y, x) relations are linear and parallel. With only one x and parallel monotonic curved relations of the types examined, linear adjustment on random samples again removes essentially all the bias if $\sigma_1^2 = \sigma_2^2$ and the distributions of x are symmetric, but may perform very erratically if σ_1^2/σ_2^2 is not near 1, or if the distributions of x are asymmetric.

Except in studies from past records, like the Cornell studies of the effectiveness of seat belts in auto accidents (Kihlberg and Robinson, 1968) matching must usually be performed before y has been measured. A drawback is the time and frustration involved in seeking matches from the available reservoirs, but this will be alleviated if computer methods like the 'nearest available' are extended to more than one x. The appeal of matching lies in the simplicity of the concept and the intuitive idea that a tight matching should work well whether the relation between y and x is linear or curved. In our studies with one x, however, the matching methods alone did not perform as well as linear regression under either a linear (y, x) relation, or a monotonic non-linear relation with $\sigma_1^2 = \sigma_2^2$ and x symmetric. Regression adjustment on matched samples also removes all the bias in the linear case and is about as effective as regression on random samples in the non-linear case. If the (y, x) relation is non-linear and σ_1^2 and σ_2^2 are very different, matching followed by regression adjustment on matched pairs performs best. Monte Carlo results on more moderate σ_1^2/σ_2^2 and asymmetric x would be helpful.

Overall, linear regression adjustment is recommended as superior to matching alone when x is continuous and only a moderate reservoir is available. In a similar comparison with more emphasis on precision, Billewicz (1965) reports that regression was more effective than matching in this respect also. However, it appears that the approach of pair matching *plus* regression adjustment on matched pairs is generally superior to either method alone.

An obvious approach not considered here is to try adjustment by a quadratic regression if this appears to fit well in both samples; there appears to be no work on this problem.

Indeed, this review has indicated numerous topics on which little or no work has been done. Even with univariate x these include research on the sizes of reservoirs needed to obtain caliper or categorical matches, on the effectiveness of the commonly used technique of incomplete matching in which members of sample 1 that lack good matches are discarded, and in methods of relaxing the restrictive assumptions of linearity and normality as suggested in Section 3.3 (and Section 5.3 for the multivariate case). For the case of a dichotomous dependent variable the only work seems to be that of McKinlay (1973).

In Sections 6.1-6.4 we have suggested several multivariate extensions of the matching methods but very little is known about their effectiveness. In this connection a survey of the commonly used methods of control, reservoir sizes and number of variables that occur in applications would be useful in guiding the scope of further research.

References

Althauser, R. P. and Rubin, D. B. (1970): The computerized construction of a matched sample. *American Journal of Sociology*, 76, 325-346.

Belson, W. A. (1956): A technique for studying the effects of a television broadcast. *Applied Statistics*, V, 195-202.

Billewicz, W. Z. (1965): The efficiency of matched samples: an empirical investigation. *Biometrics*, 21, 623-643.

COCHRAN, W. G. (1968a) : The effectiveness of adjustment by sub-classification in removing bias in observational studies. *Biometrics*, **24**, 295-313.

——— (1968b) : Errors of measurement in statistics. *Technometrics*, **10**, 637-666.

——— (1970) : Some effects of errors of measurement on linear regression. *Proceedings of the 6th Berkeley Symposium*, **I**, 527-539.

COX, D. R. (1957) : Note on grouping. *Journal of American Statistical Association*, **52**, 543-547.

DEMPSTER, A. P. (1969) : *Elements of Continuous Multivariate Analysis*, Addison Wesley.

FISHER, R. A. (1925) : *Statistical Methods for Research Workers*, 1st edition, Oliver and Boyd.

GREENBERG, B. G. (1953) : The use of covariance and balancing in analytical surveys. *American Journal of Public Health*, **43**, 692-699.

KENDALL, M. G. and BUCKLAND, W. R. (1971) : *A Dictionary of Statistical Terms*, Oliver and Boyd.

KIHLBERG, J. K. and ROBINSON, S. J. (1968) : Seat belt use and injury patterns in automobile accidents. *Cornell Aeronautical Laboratory Report* No. VJ-1823-R30.

LIGHT, R. J., MOSTELLER, F. and WINOKUR, H. S. (1971) : Using controlled field studies to improve public policy. *Federal Statistics* (report of the President's Commission) **11**, 367-402.

LINDLEY, D. V. (1947) : Regression lines and the linear functional relationship. *Journal of the Royal Statistical Society*, **B**, **9**, 218-224.

MAHALANOBIS, P. C. (1927) : Analysis of race mixture in Bengal. *J. Asiatic Society of Bengal*, **23**, 301-333.

McKINLAY, S. J. (1973) : An assessment of the relative effectiveness of several measures of association in removing bias from a comparison of qualitative variables. As yet unpublished

OGAWA, J. (1951) : Contributions to the theory of systematic statistics. *Osaka Mathematical Journal*, **4**, 175-213.

PETERS, C. C. (1941) : A method of matching groups for experiment with no loss of population. *Journal of Educational Research*, **34**, 606-612.

RAO, C. R. (1973) : *Linear Statistical Inference and Its Applications*, 2nd edition, Wiley.

RUBIN, D. B. (1970) : The Use of Matched Sampling and Regression Adjustment in Observational Studies. (Ph.D. thesis, Harvard University.)

——— (1973a) : Matching to remove bias in observational studies. *Biometrics*, **29**, 159-183.

——— (1973b) : The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, **29**, 185-203.

——— (1973c) : Multivariate matching methods that are equal percent bias reducing : Some analytic results. Unpublished manuscript.

Paper received : May, 1973.